“两微”数据查询系统实战浅析

王雪京
（中央广播电视总台，北京 100020）

摘　要：当今社交媒体已成为大数据分析的重要来源，如何在海量的媒体数据中查询到自己想要的内容成为目前主要的研究方向。本篇文章通过对微信、微博媒体大数据的查询工具的实战问题分析，并提供解决方案，为后续的用户和设计者们提供重要的参考价值。

关键词：两微系统；数据查询；实战问题分析；微信；微博；大数据　　**中图分类号：**G252.1　　**文献标识码：**A
文章编号：1671-0134（2021）02-072-04　　**DOI：**10.19483/j.cnki.11-4653/n.2021.02.020
本文著录格式：王雪京．“两微”数据查询系统实战浅析 [J]. 中国传媒科技，2021（02）：72-74+104.

1. 背景

在云计算和大数据背景下，数据呈现爆炸式增长趋势。新媒体的快速崛起，既是机遇也是挑战。媒体要实现智能匹配、融合发展，就要充分利用各种先进技术手段，充分挖掘数据价值，辅助节目生产，创新传播形式。[1] 尤其是社交媒体的话题度及对舆论的引导对于主流媒体在市场评估与收视分析中占有不可忽视的地位，通过查询分析微信、微博媒体数据，有利于引导电视台进入良性的数据生态循环，充分发挥数据自身的价值，挖掘更多的数据分析结果，提升未来业务的灵活性。[2]

为更好地分析“两微”系统的数据，即微信和微博数据，通过接入第三方清博提供的“两微”数据以及电视台自建的大数据平台自行采集的“两微”数据，实现数据查询与分析服务，[3] 为进一步促进电视台构建自主可控的自有新媒体旗舰平台，推动电视台融合发展的自有品牌发展、自有平台建设以及用户积累，是确立电视台进一步迈进新媒体大数据时代的里程碑。

在构建及运行两微系统时，面临着一些超出需求和规划的问题暴露出来，本论文通过对这些实际的案例问题进行分析并提供实际的解决方案，为后续的设计者和用户提供一些指导性的参考。

2.“两微”系统架构及数据源

2.1　系统架构

如图 1 所示，为基于 SPARK 的“两微”系统的主要设计架构。[4]“两微”数据查询系统是接入第三方清博数据及大数据平台方提供的微信和微博数据，解析导入到 kafka 队列中，通过 spark 程序消费 kafka 消息并清洗治理后导入到 ES 中。最终通过 JAVAweb 访问查询数据。主要的数据流程如下：

接入多种数据来源，有清博的微信、微博数据，和大数据平台的微信、微博数据。大数据平台方的上游是接入微博官方提供的数据和微信官方提供的数据。

消息队列接收到数据。

SPARK 应用监听消息后，存储到 HBASE。

SPARK 应用将数据存储到 ES 中。

SPARK 应用分析数据，并将分析后的结果也保存在 ES 中。

可查询微信、微博的文章数据，可查询微信、微博的账号数据。

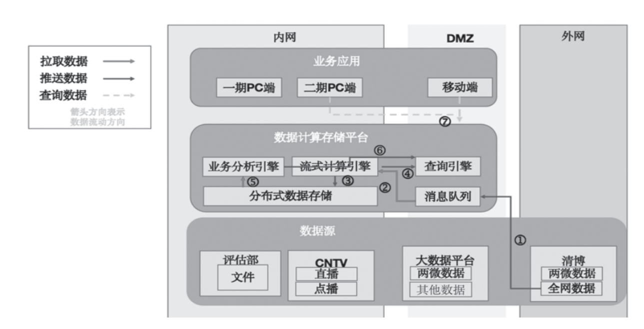


图 1　两微系统整体架构图

2.2　数据源

“两微”系统采取多数据源模式，由清博提供的微信、微博数据和大数据平台的微信、微博数据组成。当查询的账号数据是官方认证时，读取大数据平台数据。当查询的账号数据是非官方认证的，读取清博数据。当大数据平台数据未及时提供时，全部读取清博数据。

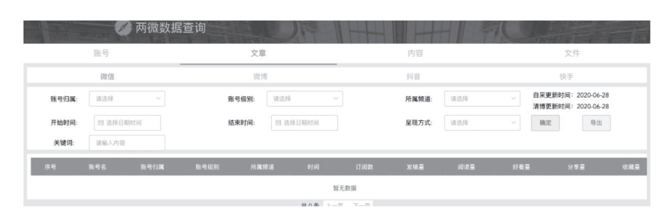
2.3　“两微”系统实际效果图



查询微信账号的指标数据



查询微博账号的指标数据



查询微信文章的指标数据



查询微博文章的指标数据



查询自定义品牌的文章数据

3. 实际案例分析

“两微”系统上线后，为台方进行市场评估及传播影响力分析方面起到了非常重要的作用。“两微”系统的特点是时效性要求很强，数据要求准确，只有这样才能提供给用户部门在进行准确的评估。但是在日常使用中，也遇到过很多不同的问题。通过与用户及上下游各系统的沟通以及对问题的分析与解决的过程，积累了很多针对“两微”数据采集与分析的经验，可以为运行维护提供有力的支撑。下面以案例的方式，将该系统运行中常见的问题进行分析与解读，可以为有“两微”数据分析需求的人提供一些浅显的帮助。

3.1 案例 1

问题表述：用户反馈，现在“两微”平台的“新媒体品牌”计算不了，输入任何新任务毫无反应。

初步问题的判定猜测：1. 用户电脑的浏览器问题，与系统不兼容；2. 网络延迟问题，导致数据无法访问；3. 后台服务发生异常，无法提供服务，前端系统缓存在客户浏览器中，所以客户依然能使用系统，但服务没反应。

初步的定位问题：首先通过不同终端的浏览器，对系统进行访问，确认问题一直存在，可以排除前两种可能。定位为问题 3。定位系统问题范围后，需要进一步查明原因，登录服务器查看服务器产生的日志。日志显示为，java.lang.OutOfMemoryError：Java heap space，时间为中午 11 点 21 分。根据时间查询 tomcat 的操作日志，操作日志显示，在查询 1 月份至 3 月份的微博文章数据。根据日志找到对应的代码，调试发现这一步是 403421 条数据载入内存中，产生共计 1.7G 的内存，服务器内存溢出。

解决问题：先通过后台写 ES 语句，查询用户想要的数据，保证用户对于数据的需求。重新启动服务，让服务正常。优化代码，发现内存溢出的语句后，进行 ES 查询语句的优化，减少中间结果数据，最终保证系统稳定运行。修改之后进行极限测试，下载多月数据，判断系统是否正常。

案例 1 的分析：系统的功能要进行极限压力测试，模拟极限查询，在极限查询条件下系统能稳定运行，即证明系统可稳定运行。为了快速方便的定位问题，要在关键的节点加入日志。

3.2 案例 2

问题表述：“今天的数据还没有更新，多久可以更新好”，用户发现数据更新时间依然是昨天。

初步问题的判定猜测：1. 上游数据源未提供数据；2. 上游数据源已经提供数据，但导入数据功能异常；3. 上游数据源提供数据，导入数据功能正常，但消息队列异常；4. 以上均正常，ES 数据库无法访问。

初步定位问题：检查 FTP 中的大数据平台中的数据是否存在。发现数据已经存在，进一步检查是否含有结束标识，发现并无结束标识，即 finished 文件。

解决问题：上游数据源未及时提供，告知上游数据提供方。待上游数据提供后，观察上游数据是否接入本系统。导入本系统成功，更新标识显示。

案例 2 分析：数据接入是有流程的，当数据无法正常接入时，优先判断数据是否接入。

3.3 案例 3

问题描述：“目前远程特别慢，似乎有问题，下载特别慢，等了 10 分钟还是没有下载成功”。

初步问题的判定猜测：1. 系统服务异常；2. 台内网络异常；3.VPN 网络异常。

初步定位问题：检查服务状态，发现服务状态正常；通过内网访问系统，响应时间正常；通过 VPN 访问网页，显示正常；模拟客户的请求，在内网访问，请求成功；通过 VPN 模拟客户的请求，请求失败。

解决问题：了解 VPN 访问机制，VPN 是通过台内代理反向映射要台内地址。VPN 有请求超时设置，超时时间为 180 秒。改变系统文件的导出机制，由同步导出，改成异步导出。系统后台自动计算，算好后供用户下载，这样不会导致 VPN 请求超时。增加超时提供。

案例 3 分析：VPN 访问时，会有一些限制，系统的设计要灵活，可以应对各种网络环境。请求异常时，要给予用户提示，否则用户不知道是系统异常还是什么原因。异常信息要清晰，分别展现出是服务异常，还是网络异常，帮助快速定位。

3.4 案例 4

问题描述：用户反馈，使用“XXXXX”作为关键词，

查无结果，如下图 2 所示。但该日期内应该能查出结果，因为该账号发表了一篇这样的文章。



图 2　案例 4 的问题描述

初步问题的判定猜测：1. 该日期内的文章（含 XXXXX 的内容的）数据未接入进系统中。2. 该日期内的文章不叫这个名字。3. 该日期内的文章还未导入至系统中。

初步定位问题：去微博查看找到用户指定的文章。通过该微博的url，搜索数据库，发现数据库中含有该文章。文章的名称也是一样的。检测数据，发现同一篇文章会多次出现在数据库中，因为要保存文章在不同时间点的状态。在数据库中查询客户搜索条件下的日期内文章，发现文章名称和最新的名称不一致，导致无法在指定日期内查询出该文章。

解决问题：基于变化，完善系统的健壮性，将微博名称的变更也考虑进去。

案例 4 分析：数据分析工具的准确性，依赖于数据源头的稳定性，但数据源头本身就是多边的，需要将系统设计的更加灵活，才能增加系统的健壮性。

3.5　案例 5

问题描述：用户反馈，“今天自采数据还未更新，是什么原因呢”。

初步问题的判定猜测：1. 上游未提供数据。2. 上游提供数据，但导入数据功能异常。3. 上游提供数据，导入数据功能正常，但消息队列异常。4. 以上均正常，ES 数据库无法访问。

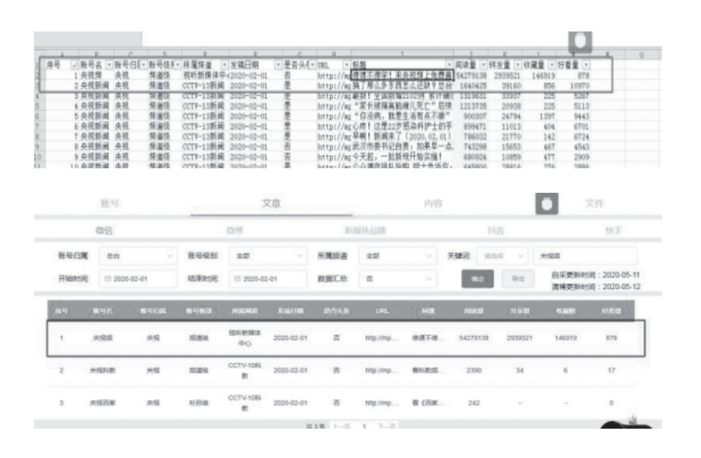
初步定位问题：上游数据已提供。发现导入数据功能异常，导入数据功能未执行。确认定时执行导入数据功能的系统用户无法启该功能，最终发现为系统用户密码超期。

解决问题：修改系统用户密码，使系统用户生效。系统要检查关键定时任务是否触发，触发后告知用户。整理系统用户密码超时文档，定期检查即将过期的用户，并改密码。

案例 5 分析：合理有效的巡检，是系统稳定运行的前提条件。保证文档的完整性，并有巡检人员及时查看。定期修改密码，保证系统的安全性。

3.6　案例 6

问题描述：“这是一个微博文章的数据，阅读量大的夸张，需要核实一下数据的准确性”。如下图所示：



初步问题的判定猜测：1. 该数据是否为上游数据源提供数据。2. 该指标是否经过计算。

初步定位问题：该数据是上游提供数据，并且该数据未经过计算。最终确认微信官方提供数据异常。

解决问题：告知上游数据源重新接入数据。手动触发，将该日的数据接入到本系统中，再次检查数据是否合理。

案例 6 分析：数据处理要包含一些快速处理异常的机制。当上游数据异常，需要系统级的重新接入时，可以快速响应，用最短的时间处理问题。

3.7　案例 7

问题描述：用户反馈，“昨日的微博的直播数据均为 0”。

初步问题的判定猜测：1. 数据库中的数据为 0；2. 上游提供的原始数据为 0；3. 上游数据异常或者上游的上游数据异常。

初步定位问题：查询本系统的数据为 0。数据源的直播数据也为 0。微博刚刚修改了机制，直播回放以后都会变成视频，所以没有直播数据了，视频数据量是基于直播数据累计的。

解决问题：调查数据源头的变更，查明后修正接受数据的方法。

案例 7 分析：定期查看数据源头的变化，在本系统中快速做出响应。

4. 总结与展望

通过对“两微”系统查询工具的实战问题分析，了解针对查询系统出现的问题，首先明确问题定位，给出初步的问题来源，其次通过查询相关工具和内容定位问题的发生，最后，提出对应的解决方案完善查询系统的功能。科学合理地描述问题、定位问题、解决问题，是保证“两微”系统查询工具的运维稳定性与可靠性，对用户和其他设计者们具有重要的参考价值。[4]

在未来工作中，从“两微”系统查询工具实战问题分析中积累发现问题、解决问题的实战开发经验，扩展到电视台其他系统，为电视台的新媒体数据搜索和数据的价值分析提供强有力的运维技术保障，[5] 稳定提升系统稳定性，强化数据查询搜索分析带来的实际效益。媒

（下转第104页）